NOVA IMS Information Management School

MDSAA

Master's Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 2: Sales Forecast

Diogo Reis, number: 20230481

Maria Almeida, number: 20230489

Marta Jesus, number: 20230464

Rita Matias, number: 20230496

Tomás Louro, number: 20230285

Group I

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

April, 2024

INDEX

1. EXECUTIVE SUMMARY

The consulting team tasked with forecasting sales adopted the CRISP-DM methodology to address the underlying problem of Siemens.

The workload started with a meticulous data understanding phase, during which a complex exploration of the data that was provided by Siemens Advanta. This exploration was complemented by an extensive industry research, which illuminated market trends and external variables potentially impacting the company sales. Through this analysis, numerous visualizations were generated to contextualize sales figures alongside macroeconomic indicators.

The next crucial phase was the preprocessing. In this phase sales data was adjusted to inflation and aggregated by month, year, and product. It is important to note that the stationarity of the different product groups was also studied, highlighting that only product 8 is non-stationary. In addition, the data regarding macroeconomic indexes was treated to fill missing values. The approach taken to impute them was informed by the Kolmogorov test.

Next, the preprocessing was continued with a feature engineering stage, where six lagged features and two rolling windows were created, so that time dependencies can be captured – features that later proved to be pivotal in sales prediction. The next step encompassed a robust feature selection phase embedding three supervised machine learning models, in which for each product group, five commonly selected features were chosen among the top 10 ranked as most significant features of each model. Afterwards, outliers were detected using z-score and rectified through interpolation method.

The end of the preprocessing phase yielded the team with three datasets for each product ready for modelling: one only considering the original features, another with both the original features and the lagged features, and a third that was built based on the first but addressed outliers. Each of these datasets was beforehand divided so that 80% is used for training and the remaining 20% for validation, laying a robust foundation for the subsequent modelling phase.

On what regards the modelling phase, it comprised the training of various models for the three datasets of each product group. At the end, and regarding the *Evaluation* step, all *Root Mean Squared Error* (RMSE) by model and product were gathered and compared between each other, the better performing model, say the one with the lowest *RMSE*, was selected to predict the sales for that same product. Thus, the combination of models and datasets performing better for each product was selected to be used for the test dataset.

To conclude, sales per product were successively forecasted for the requested period with a strong belief in the quality of the obtained predictions.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. BUSINESS BACKGROUND

Siemens Advanta, multinational company, based in Munich, Germany, is composed by three distinct branches of business – Industrial Business, Services and Other Business. Part of the Industrial Business branch is the Smart Infrastructure Division, which besides other activities, produces smart power distribution components. The production range of this business unit goes from Medium-voltage Components, Systems & Solutions, to Low-voltage Systems and Automation, Protection & Communication for high and medium-voltage systems. In a fast-paced world it is essential to forecast the sales accurately, taking advantage of the digital tools at our disposal.

2.2. BUSINESS OBJECTIVES

The development of an *Artificial Intelligence-driven* sales forecasting model helps companies overcoming several problems and difficulties, such as the unconscious bias introduced by the judgement of multiple stakeholders or the high demand for resources that this task may require. Additionally, a fine-tuned model for sales forecasting also enables the company to better adjust its supplying needs and production plans for the different products to match future sales, achieving a more efficient overall operation.

2.3. BUSINESS SUCCESS CRITERIA

The success of this process can be assessed through the following criteria: accurate monthly sales forecast of the selected product groups of a specific business unit in Germany. This will enable an improved inventory management, enhanced production planning, mitigation of risks associated with economic fluctuations and optimization of marketing and sales efforts. These criteria can be assessed by the board of directors of Siemens in collaboration with the project team.

2.4. SITUATION ASSESSMENT

In order to enable the sales forecast for Siemens Advanta specified groups, it was necessary to collect data from historical sales. This phase was facilitated by Siemens' provision of a database, which not only included sales data suitable for training and testing the predictive models, but also encompassed macroeconomic indicators that could potentially influence sales trends. Siemens demonstrated a robust technological infrastructure, therefore having advanced software systems capable of recording historical sales data, as well as the in-house expertise in data management. To leverage the data to predict sales, it was required a team of data science experts with knowledge in Machine Learning and with deep comprehension in Time Series, which is where Group I play a role.

Further on, it is possible for the sales department of Siemens Advanta to make data-driven decisions with the goal of optimizing future sales strategies.

One of the major challenges faced during the project was the limited timeframe given for its development, since it was given three weeks to complete all the steps as better as possible. Additionally, the lack of communication from the stakeholders might have resulted in some misunderstandings since it was not possible to clarify some details.

2.5. DETERMINE DATA MINING GOALS

The goal of this project is to take advantage from the sales data between October 2018 and April 2022 together with important macro-economic indicators, such as Consumer Price Index, to predict sales for the period of May 2022 to February 2023.

The primary goal of this predictive model is to achieve performance on the validation set, which in this specific case will be assessed through a minimization of the *Root Mean Square Error* (RMSE). Subsequently, the objective is that the model delivers accurate predictions for the test set, aligning closely with the actual data upon its availability. Success relies on the robustness and validity of the methodologies applied throughout data preparation, pre-processing, and modelling. To achieve this, a commitment was made to leverage expertise in *Data Science* to best select and apply the most effective methods used for instance when handling missing values, outliers, feature engineering, and feature selection, as well as optimizing model parameters through grid search.

3. METHODOLOGY

3.1. DATA UNDERSTANDING

Comprehending the dataset stands as an essential step within every project based on the CRISP-DM methodology. For this reason, heavy investment was made in this step to ensure that the data from both main datasets, *Sales* and *Market*, were deeply explored and descripted ensuring quality through the process.

3.1.1. Sales dataset

The *Sales* dataset – *Case2_Sales data* – comprises 9802 records of historical sales in euros on a daily periodicity, from October 2018 to April 2022. For each record it maps the GCK of the product, therefore characterizing its group product according to the smart power distribution from medium-voltage to low-voltage, confirming that 14 different group products had sales records. To proceed with the understanding phase, data types were checked and handled to better align with their analytical purpose and to enhance the efficiency of the models. It was then verified the absence of duplicated entries and sales records with null value were removed, thus resulting in a dataset with 2668 records. These records were then aggregated by the same month and year, and by the GCK of the product.

To better understand the evolution of sales throughout the considered time, various visualizations are depicted. *Figure 1* exhibits a volatile behavior of sales throughout time, with numerous peaks and troughs. Although the data does not show a clear long-term trend, it does show significant fluctuations that could be due to external factors or events specific to the industry. For instance, the dip in 2018 is possibly connected to changes in investment approaches within the electric power industry, as suggested by a McKinsey study *[1]*. Additionally, the drop in sales in 2021 might be attributed to the economic impact of the global COVID-19 pandemic.

A closer examination of the total sales on a quarterly basis reveals a consistent annual pattern (*Figure 2*). Each year tends to start with the lowest sales in the first quarter, which is then followed by a rise in sales during the second quarter. The peak sales period is consistently observed in the third quarter, with the fourth quarter being the second highest. This trend likely reflects a seasonal increase in sales

towards the end of the year, which could be attributed to the industry's typical year-end purchasing trends. While this sales pattern is consistent, there are slight variations from year to year.

In terms of sales distribution among product groups *(Figure 3-4),* GCK 1 dominates, accounting for approximately 60% of total sales, GCK 3 comes in second, contributing 20%, followed by GCK 5, which makes up roughly 16%. All other product groups combined total less than 7% of overall sales.

To gain a detailed comprehension of the sales trajectory for each product group, an in-depth analysis was performed, encompassing numerous detailed visualizations to aid in the interpretation of the data (*Figure 5* and *Figures 6-7*). *Visualization number 5* makes it apparent that there is an overall consistency of sales pattern across the different product groups throughout the considered period.

Furthermore, an examination of the year-on-year and month-by-month sales progression for each product group reveals certain periods where product sales trends align, therefore underscoring a parallelism in consumer purchasing habits during specific months across multiple years. *(Figures 6-7)*

3.1.2. Market dataset

The *Market* dataset – *Case2_Market* – comprises data on important macro-economic indices for Siemens in its most important countries from February 2004 to April 2022. These indexes have as their base year 2010, for that reason index value is 100. Detailed information on these indexes and for which countries each of them is present can be found in Text 1.

To move forward with the understanding phase, data types were checked and handled to best suit their analytical purpose, and it was also checked that there were no duplicate entries.

The next stage to have a deeper insight on these indexes was to plot numerous visualizations. The first one was the *Production vs Shipment* index by country to understand how these two fluctuate together over time (*Figure 8*). This analysis revealed that in some countries, such as Japan and China, these indexes are almost the same throughout the considered time. This may be due to the existence of strict policies that closely align production output with market demand, therefore promoting a highly synchronized economy in which what is produced is quickly shipped. Conversely, countries like France and the United Kingdom, although production and shipment trends show similar patterns, they do so at different magnitudes. This discrepancy may be attributed to several factors, such as periods of high production to build up inventory before peak demand seasons, which is later reflected in shipment increases.

Subsequently, the evolution of world prices of raw materials was analyzed (*Figure 9)* and showed to have high volatility and some inter-commodity correlation. It was noticed that significant peaks in the overall prices but more pronounced for natural gas are observed during 2008 and 2022, corresponding to periods of financial recession, caused by the Lehman Brothers collapse and COVID-19, respectively. The recovery period after the 2008 financial crisis is relatively stable until around 2020 when prices escalate due to the pandemic hit.

Regarding *Producer Prices - Electrical Equipment* (*Figure 10*), it was observed that generally the different countries have their own behavior, however they show a similar behavior for the period after the 2008 financial crisis and after the COVID-19 in where all of them increase. Lastly, concerning

Production Index - Machinery and Equipment (*Figure 11*) and *Production Index - Electrical Equipment* (*Figure 12*) both indexes disclose an identical evolution, with clear seasonality over the years.

3.2. DATA PREPARATION

On data preparation and pre-processing different techniques were used to clean and transform the data to be used on the following phase which is modelling. The better this step is done the better performance and capacity to generalize the models will get.

3.2.1. Sales dataset

The preparation of the *Sales* dataset started with adjusting the sales values to inflation throughout the considered time, as this would allow to have a more realistic picture of sales and more accurate comparisons of sales performance over time. To do so, the *Consumer Price Index* based on 2020 was extracted from *DEStatis* [2]. This adjustment was made multiplying the total sales of each product per month by the corresponding CPI value for that month. Having this adjustment, total sales for each product group were graphically analyzed again.

The next step was to understand if each product was stationary or not, meaning that the series properties such as mean, variance and autocorrelation structure do not change over time if it is stationary. This assessment is crucial as it stands as an assumption for numerous forecasting models, such as *ARIMA*. To perform this evaluation, lag plots for each product were visualized and stationary was inferred as they did not seem to show any pattern. To ensure the validity of these results, the *Augmented Dickey-Fuller* test was performed, confirming the above results with exception for product group 8 that given the p-value greater than 0.05 is characterized as non-stationary. Further assessment for GCK 8 was done to get insights on its trend and seasonality. After further graphical exploration, and as can be seen on Figure 33 there is a clear seasonality related to the evolution of sales regarding GCK 8.

3.2.2. Market dataset

The preparation of the *Market* dataset started with the missing values check and its treatment. The first step was to separate the variables with missing values in a separate data frame and then associate the date corresponding to the missing record. To have the most accurate value to fill missing values, the team preformed a *Kolmogorov-Smirnov* test, to check if the variable in cause followed a normal distribution, then depending on the outcome, performed one of the three following paths. If the index follows a normal distribution, as *MAB_ELE_SHP840* does, the missing values were inputted with a prediction based on a *Simple Moving Average*. Regarding indexes that had a significant number of missing data, such as *MAB_ELE_SHP826*, the fulfillment of the missing values was based on a normal distribution. If the variable does not follow a normal distribution, as it is the case of *PRI27826_org*, the correction of missing values was done by performing a season decomposition and fitting the *Exponential Smoothing Model* to that series and finally forecast the future missing values. Lastly, if the missing values were from periods of time before 2018 the group opted not to handle them, since the target predictions only start in 2018.

3.2.3. Market_Sales

After preparing both *Sales* and *Market* datasets, the next step was to merge the two resulting datasets in one and prepare it to the modeling phase of the project. Firstly, a check to ensure that all products

had sales in all considered months was performed. The outcome of this check indicated that products 9, 14 and 20 had months where no sales were logged, those missing records were inputted using a linear interpolate method. After ensuring that no data was missing, the *Sales* and *Market* data frames were merged using as merging key each product unique identifier, storing as its values the indexes and the sales value by month and year.

Following that, an approach regarding the creation of *Lag Features* was taken. Lag features consider previous values of the target variable or other relevant variables at different time lags. This way, by selecting the appropriate lagged features temporal dependencies can be captured. Initially six lagged features were created along with two rolling windows. These lagged variables take as their value for a given month the value of that same variable from the previous month, up to a maximum of 6 months ago, respectively. For example, for the month of April, lag 1 takes the value of sales of March, lag 2 takes sales of February, and successively. Concerning the created rolling windows, the first computes the average of the three past months and the second computes the average of the six past months. From here two datasets were obtained, the original and another including lag features.

Before continuing to the feature selection phase, it was crucial to perform the train-validation split. This split was done considering the chronologicity of the data, therefore the training set corresponds to the first 80% of the months and the validation corresponds to the last 20% of the given months.

Finally, an in-depth methodology was taken in the feature selection phase. To start, the *spearman method* was used to check correlations between each input feature and the target for each product group and it was verified that there were no significant correlations (all absolute values below 0.75). Additionally, the same method was used to check correlations between input features for each product, and variables with a correlation higher than 0.89 were dropped since they are assumed to be redundant, resulting in a data frame with 27 features. The same approach was taken for the dataset including lagged features, where product 20 *Sales_CPI_€_rolling_mean_3* exhibited a correlation with the target variable above 0.75, indeed later this variable showed to have significant importance. For the correlations between Index features an unchanged approach was performed, causing the removal of several indexes resulting in a dataset with 35 features for all products, except for product 4 and 12 where the resulting dataset comprehended 34 features.

Following the steps above, an ensemble feature selection was performed, using *Random Forest*, *Gradient Boosting* and *XGBoost*. This is a robust technique since it identifies features that are consistently important across models. Moreover, it mitigates the risk of choosing features that are only relevant due to the specific biases of single model. Concerning the three models used for ranking feature importance, it was decided that only the top 5 commonly ranked should be kept for each product based on their occurrences among the top features obtained from each performed model. This process was applied to both the original dataset and the one that includes the lagged features. Finally, the selected features regarding each product were stored in separate datasets, for both original and lagged dataframes.

A *Z-Score* approach was used to detect potential outliers as it is an efficient statistical method for identifying data points that are unusually far from the mean. This method assumes that the data is normally distributed. This resulted in a total detection of seven outliers across the dataframes, which were promptly corrected using an interpolation method.

This stage of the CRISP-DM methodology, which sets the grounding to the modeling phase, ends with three distinct datasets for each product – the original, the lagged and the one that includes the corrected outliers.

3.3. MODELING

As the grounding is set for the modelling the process of selecting the best model started. Through this phase of the project, the main goal was to find between Supervised Learning, Prophet, and Large Language Models, which of those have the better capability to generalize predictions for each product sales, meaning which perform better on never-before-seen data.

Regarding Supervised Learning models, *XGBoost* was the chosen model to predict the sales values. Firstly, a grid search was run in order to find the best parameters, however, so as to overcome computational power barriers only some parameters were tested – *n_estimators*, used to select the number of boosted trees; *max_depth*, that represents the maximum depth of each tree; *learning_rate*. [8]

Concerning Prophet model, which is mainly used for time-series forecasting that takes as inputs two columns – data stamp, ds, and the target variable, y. Prophet has also properties to work with multiple seasonality's, changing growth rates and the ability to models special days. The initial step was to rename the columns *Date* and *Sales_CPI_€* and then the logic behind them was the same as used on supervised learning models. [9]

With respect to *Large Language Models, TimeGPT* was used, which is a recent model that has been pre-trained to generate forecasts for time series data enabling accurate predictions, taking as input only historical values without training. It looks at a window of past data and predicts based on patterns found and extrapolations into the future data. The first step of this algorithm is a token activation, then the rationale applied was analogue to the other models, it is iterating over each dataframe applying the model to get predictions and calculate the evaluation metric. Then a comparation of all dataframes per product was made so that was possible to find the one that achieved the best results. The model was also run with a monthly frequency and using the renamed columns ds and y. [10]

The following phase was to predict all the index values for the new dates in order to, after this step, be possible to make the final predictions. Firstly, both remerged_data and test dataframes were merged and then the approach used was to check which features by product follow or not a normal distribution. For the ones that follow, the predictions were made based on the mean and the standard deviation. For the ones that didn't follow a normal distribution, were checked if they were stationary or not. The ones that were predicted using simple exponential smoothing – the method used for data that doesn't present any seasonality or trend - and the not-stationary ones were predicted using the hold winters method. Both models are from the Error – Trend – Seasonality model groups.

3.4. EVALUATION

Evaluation is a critical phase, it is where further assessment to the quality and validity of the resulting sales forecast is made. Inside *sklearn* library are several metrics that can be used to evaluate regression problems, such as Explained Variance, *Mean Absolute Error* (MAE) and *RMSE*. The metric chosen to evaluate the performance of the models was the *RMSE*, this metric computes the square root of the expected value of the squared error. During the training process, a function to calculate the *RMSE* for

all dataframes and models was created and frequently applied. After all, *RMSE* by model and product were gathered and compared between each other, the better performing model, say the one with the lowest *RMSE*, was selected to predict the sales for that same product. *[11] [12]*

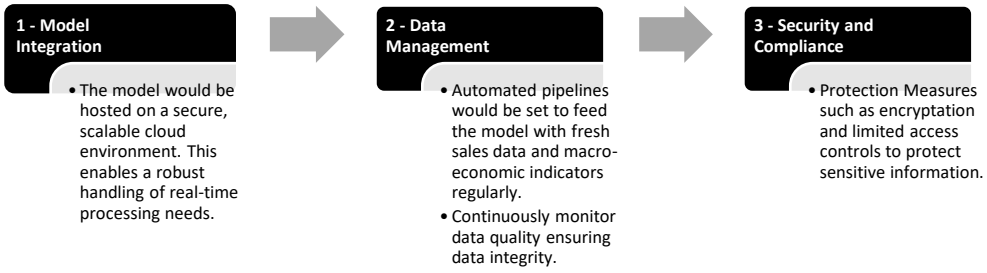
4. RESULTS EVALUATION

Considering the chosen metric to assess the quality of the performance of the models, the following table is presented, accommodating a summary regarding each product and the model used to predict it's sales.

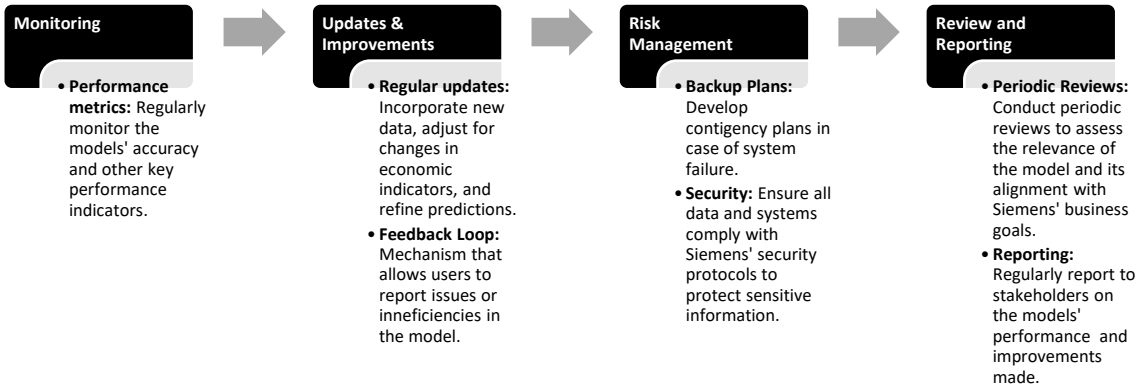
Product ID	Choosen Model	Dataframe used	Best RMSE	Best Parameters (if applicable)
1	Prophet	original	4436100	
3	Prophet	outlier	2018385	
4	XGBoost	lag	115017	learning_rate = 0.05, max_depth = 3, n_estimators = 1000
5	TimeGPT	lag	3242273	
6	XGBoost	original	207477	learning_rate = 0.01, max_depth = 5, n_estimators = 1000
8	Prophet	lag	610963	
9	XGBoost	lag	4242	learning_rate = 0.01, max_depth = 5, n_estimators = 100
11	Prophet	original	780347	
12	XGBoost	lag	106137	learning_rate = 0.01, max_depth = 5, n_estimators = 500
13	XGBoost	original	12410	learning_rate = 0.01, max_depth = 5, n_estimators = 100
14	TimeGPT	original	15650	
16	TimeGPT	original	89173	
20	TimeGPT	original	2250	
36	XGBoost	outliers	15602	learning_rate = 0.05, max_depth = 5, n_estimators = 100

5. DEPLOYMENT AND MAINTENANCE PLANS

The successful prediction of sales marked a significant milestone of the project in hands. Thus, it is imperative to outline a deployment plan to ensure the integration of the developed forecasting model into Siemens' operational framework.



On what regards the suggested maintenance plan, it englobes four phases. These stages would be performed by a supporting team.



6. CONCLUSION

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

In the pursuit for sales forecasting accuracy, it is acknowledged the necessity for continual model enhancement. This could be done through the incorporation of the following points:

- Having access to more years of records from previous sales.
- Experiment alternative approaches to impute missing values.
- The corrective measure applied to outliers in the original dataset proved beneficial for some product groups, therefore it would be interesting to extend this corrective technique to the dataset including lagged features. Additionally, different techniques to handle outliers could be tried.
- Experiment an increased array of models, as broadening the modeling spectrum increases the likelihood of discovering more potent predictive relationships within the data.
- Experiment forecast combinations, that is, averaging the resulting forecasts of different models on the same time series.

By embracing these considerations for model improvement, the goal is to refine the forecasting capabilities.

To finish, next steps would involve incorporating these considerations for model improvement along with implementing the above-mentioned deployment and maintenance plans. Moreover, to complement, the team suggests an adoption of continuously predicting sales both in the short and long term, while being able to adapt and incorporate recent information into predictions, as it becomes available – a commonly used approach called Walk Forward technique.

7. REFERENCES

[1] *Electric grids are evolving | McKinsey.* (n.d.). Www.mckinsey.com. https://www.mckinsey.com/industries/electric-power-and-natural-gas/our-insights/modernizing-the-investment-approach-for-electric-grids

[2] *Federal Statistical Office Germany - GENESIS-Online.* (2024, April 11). Www-Genesis.destatis.de. https://www-genesis.destatis.de/genesis//online?operation=table&code=61111-0002&bypass=true&levelindex=0&levelid=1712761640641#abreadcrumb

[3] Lazzeri, F. (2020). *Machine Learning for Time Series Forecasting with Python*. Wiley

[4] Brownlee, J. (2016, December 13). *Basic Feature Engineering With Time Series Data in Python.* Machine Learning Mastery. https://machinelearningmastery.com/basic-feature-engineering-time-series-data-python/

[5] Brownlee, J. (2017, March 28). Feature Selection for Time Series Forecasting with Python - MachineLearningMastery.com. *MachineLearningMastery.com.* https://machinelearningmastery.com/feature-selection-time-series-forecasting-python/

[6] *Forecasting.* (n.d.). Open Time Series. https://opentimeseries.com/python_packages/forecasting/

[7]*Basic Time Series Analysis & Feature Selection.* (n.d.). Kaggle.com. Retrieved April 10, 2024, from https://www.kaggle.com/code/creatrol/basic-time-series-analysis-feature-selection

[8] *Python API Reference — xgboost 1.6.1 documentation.* (n.d.). Xgboost.readthedocs.io. https://xgboost.readthedocs.io/en/stable/python/python_api.html#module-xgboost.sklearn

[9] *Quick Start*. (2017). Prophet. https://facebook.github.io/prophet/docs/quick_start.html

[10] *TimeGPT Quickstart.* (n.d.). Time GPT (Beta). Retrieved April 10, 2024, from https://docs.nixtla.io/docs/timegpt_quickstart

[11] *sklearn.metrics.mean_squared_error — scikit-learn 0.24.2 documentation.* (n.d.). Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html

[12] *3.3. Metrics and scoring: quantifying the quality of predictions*. (n.d.). Scikit-Learn. https://scikit-learn.org/stable/modules/model_evaluation.html#mean-squared-error

[13] *Scoring a time series model | IBM Cloud Pak for Data as a Service.* (n.d.). Dataplatform.cloud.ibm.com. Retrieved April 10, 2024, from https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/autoai-ts-score.html?context=cpdaas

8. APPENDIX

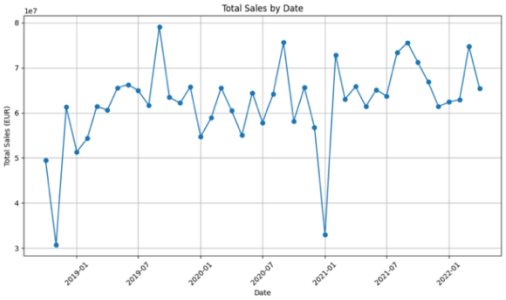


Figure 1- Total Sales by Date

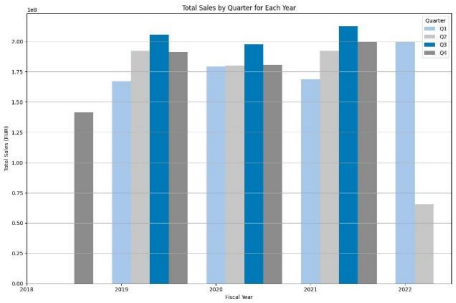


Figure 2- Total Sales by Quarter for each Product

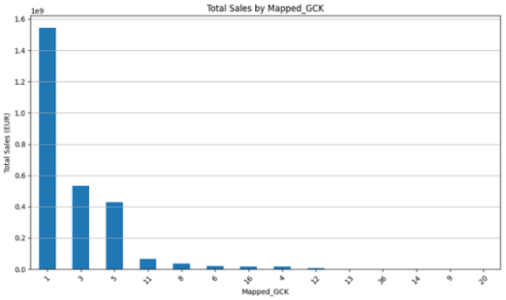


Figure 3- Total Sales by product

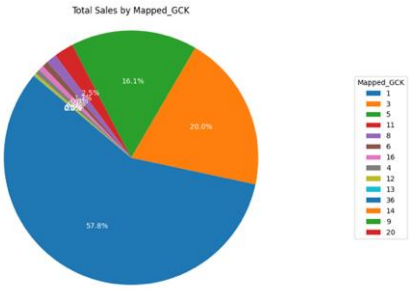


Figure 4- Pie Chart of total sales by product

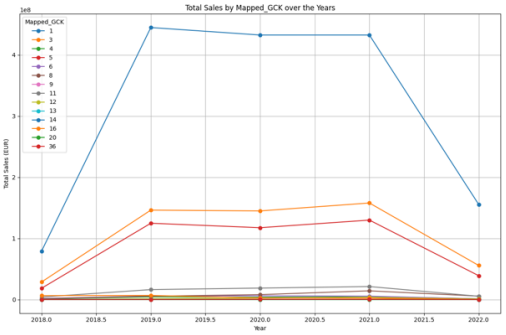


Figure 5- Total Sales by product

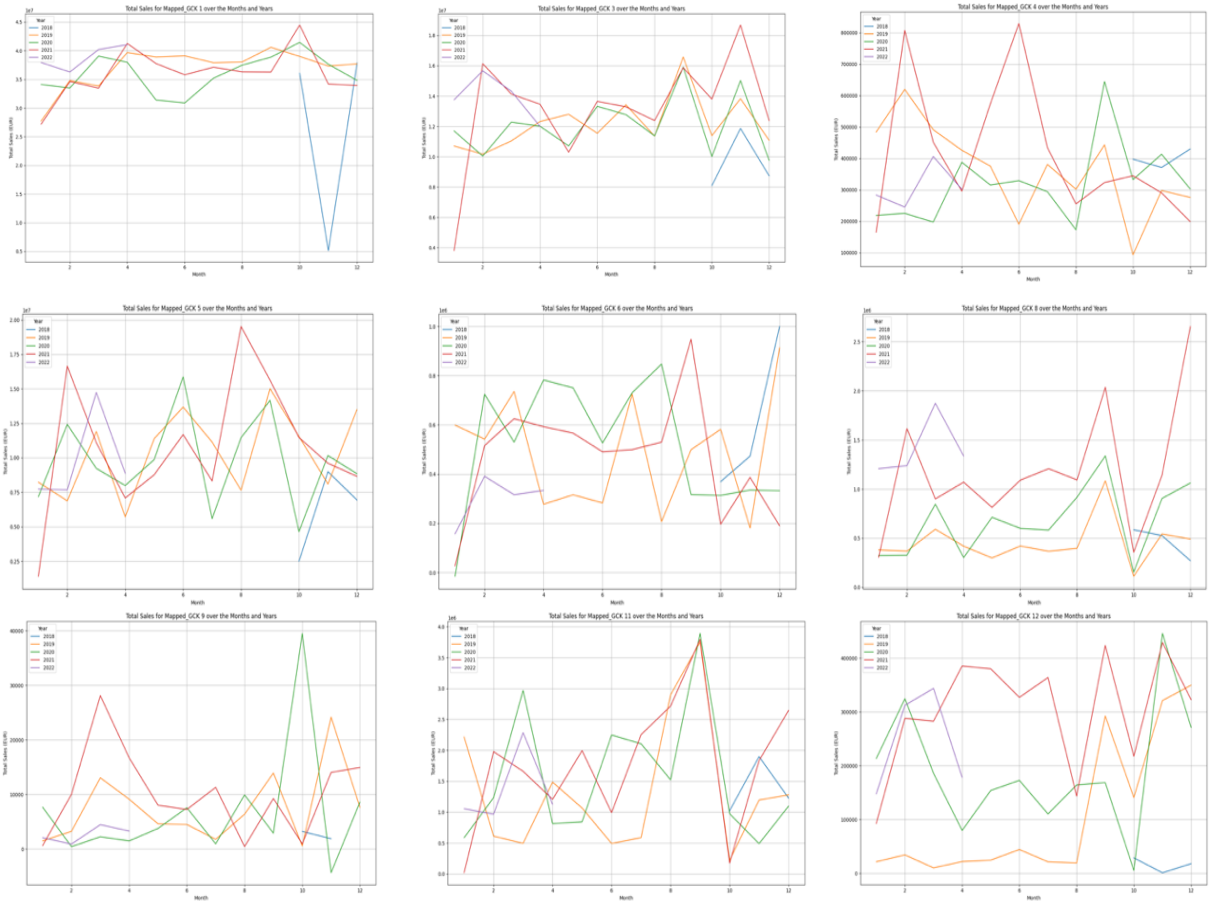


Figure 6- Total Sales for each product by Months and Years

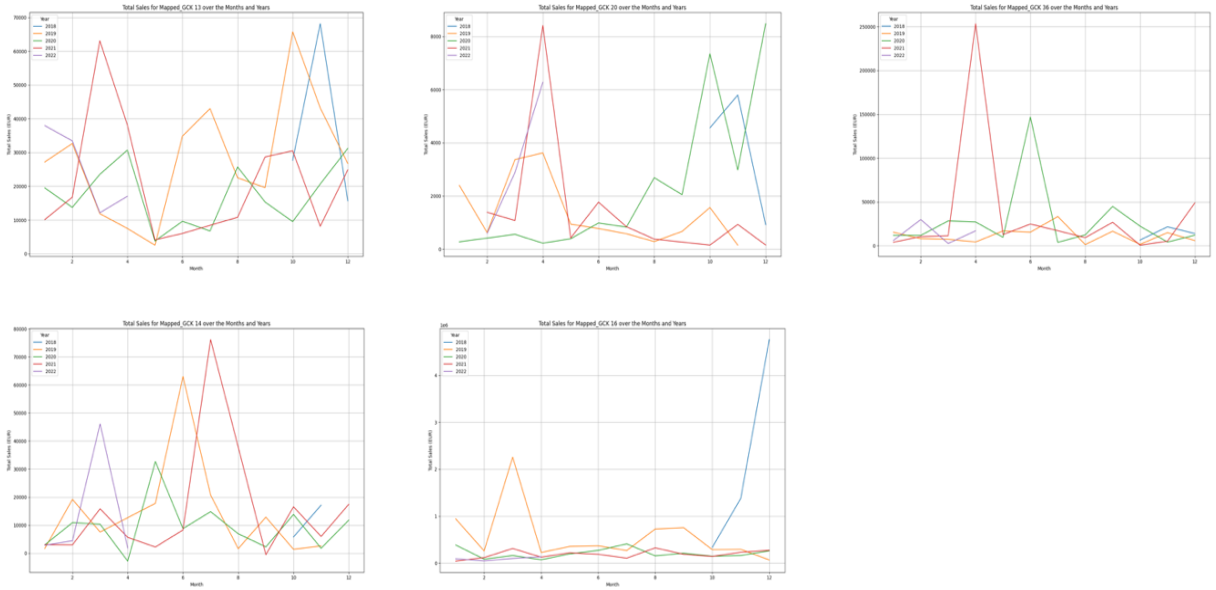


Figure 7- Total Sales for each product by Months and Years

Text 1

The 2 following indexes are presented for China, France, Germany, Italy, Japan, Switzerland, United Kingdom, United States, and Europe:

- Production Index Machinery & Electricals
- Shipments Index Machinery and Electricals

The 6 following indexes refer to Raw Materials World Prices:

- Base Metals
- Energy
- Metals and Minerals
- Natural Gas
- Crude Oil, average
- Copper

The following index is presented for United States, United Kingdom, Italy, France, Germany, and China:

- Producer Prices Electrical equipment

The 2 following indexes are presented for United States, World, Switzerland, United Kingdom, Italy, Japan, France, Germany:

- Production Index Machinery & Equipment n.e.c.
- Production Index Electrical Equipment

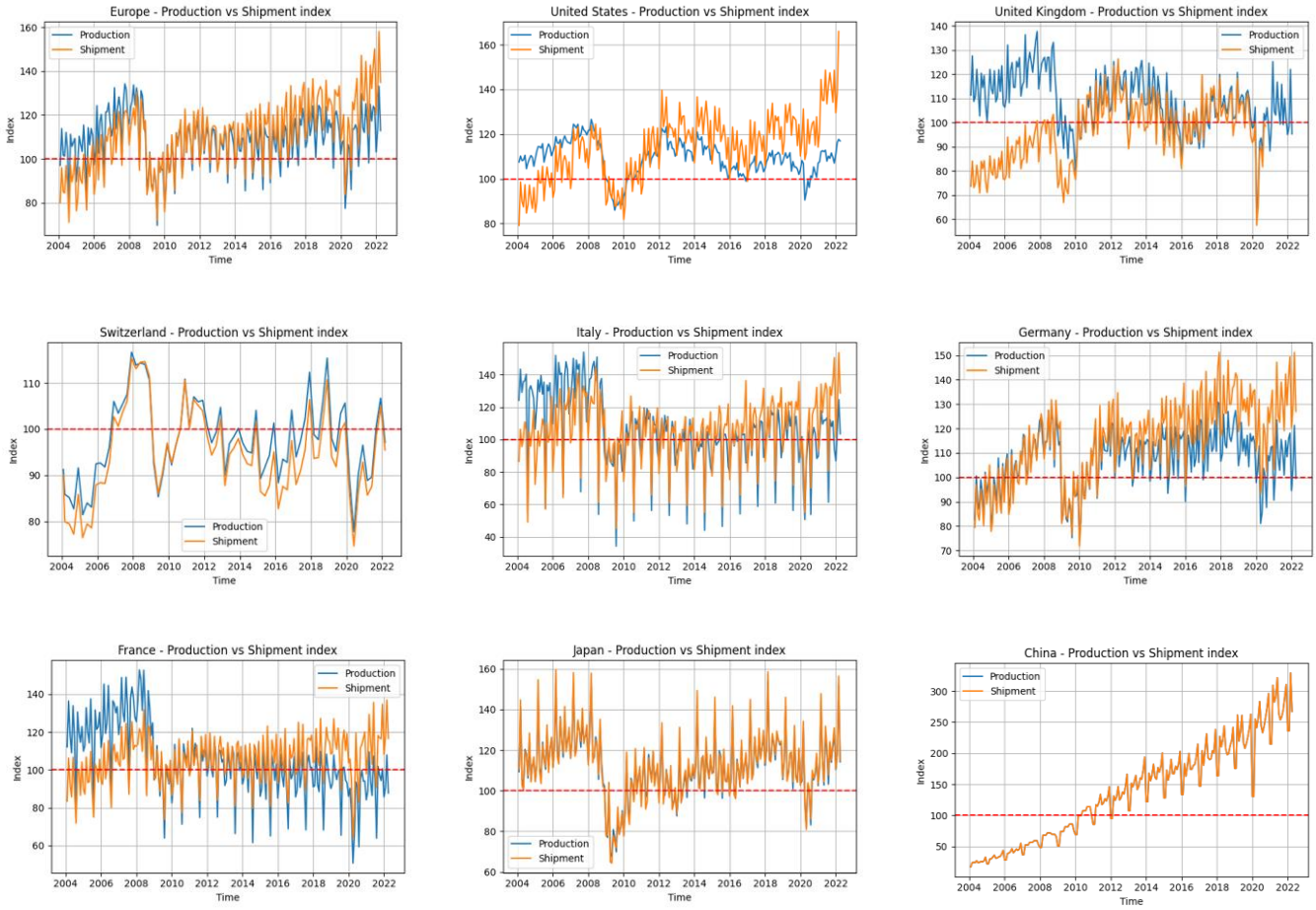


Figure 8- Shipment and Production Indexes

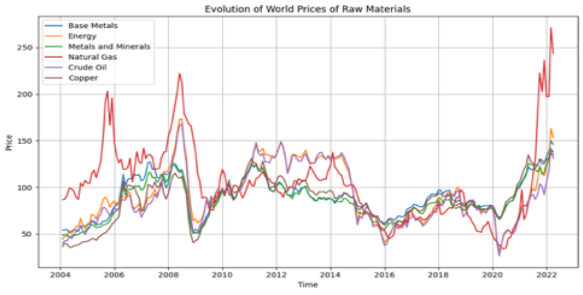


Figure 9- World Prices

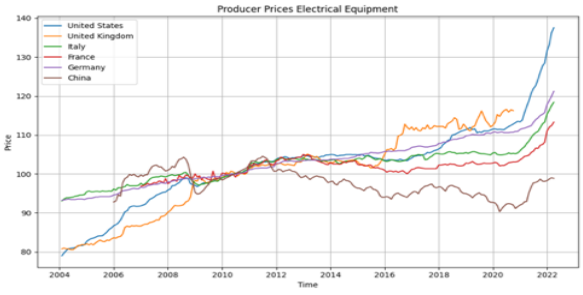


Figure 10- Producer Prices

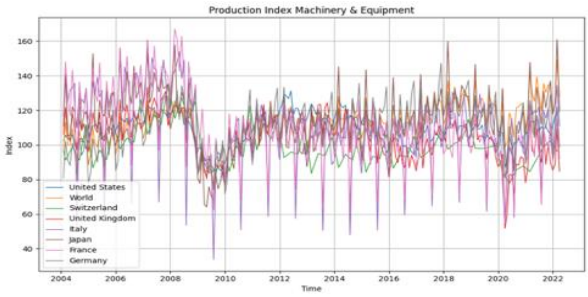


Figure 11- Production Index Machinery and Equipment

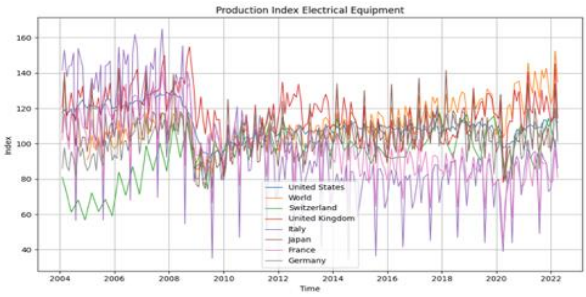


Figure 12- Production Index Electrical Equipment

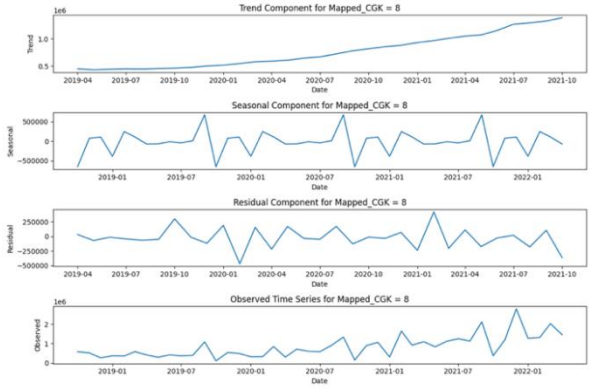


Figure 13- Product 8

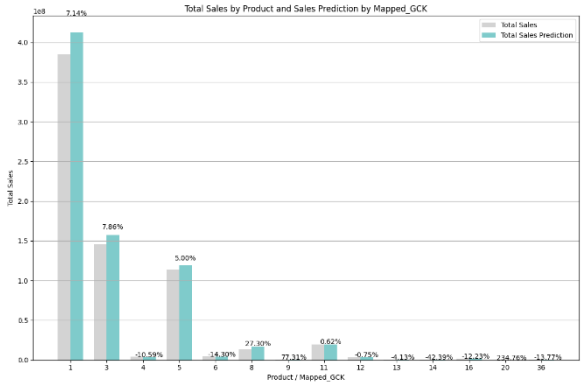


Figure 14- Total Sales and Predictions by Products